Published online 3 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D865–D870***
doi:10.1093/nar/gks1025

Xenbase: expansion and updates of the *Xenopus* model organism database

Christina James-Zorn[1], Virgilio G. Ponferrada[1], Chris J. Jarabek[2], Kevin A. Burns[1], Erik J. Segerdell[3], Jacqueline Lee[2], Kevin Snyder[2], Bishnu Bhattacharyya[2], J. Brad Karpinka[2], Joshua Fortriede[1], Jeff B. Bowes[2], Aaron M. Zorn[1,*] and Peter D. Vize[2,*]

[1]Division of Developmental Biology, Cincinnati Children's Research Foundation, 3333 Burnet Avenue, S3.620, Cincinnati, OH 45229-3039, USA, [2]Department of Computer Science, University of Calgary, 2500 University Drive NW, Calgary, AB T2N1N4, Canada and [3]Oregon Health & Science University, Ontology Development Group, Portland, OR 97239, USA

Received September 14, 2012; Revised October 2, 2012; Accepted October 3, 2012

ABSTRACT

Xenbase (http://www.xenbase.org) is a model organism database that provides genomic, molecular, cellular and developmental biology content to biomedical researchers working with the frog, *Xenopus* and *Xenopus* data to workers using other model organisms. As an amphibian *Xenopus* serves as a useful evolutionary bridge between invertebrates and more complex vertebrates such as birds and mammals. Xenbase content is collated from a variety of external sources using automated and semi-automated pipelines then processed via a combination of automated and manual annotation. A link-matching system allows for the wide variety of synonyms used to describe biological data on unique features, such as a gene or an anatomical entity, to be used by the database in an equivalent manner. Recent updates to the database include the *Xenopus laevis* genome, a new *Xenopus tropicalis* genome build, epigenomic data, collections of RNA and protein sequences associated with genes, more powerful gene expression searches, a community and curated wiki, an extensive set of manually annotated gene expression patterns and a new database module that contains data on over 700 antibodies that are useful for exploring *Xenopus* cell and developmental biology.

INTRODUCTION

The *Xenopus* embryo serves as a powerful *in vivo* model to explore the basic mechanisms of cellular function. The embryos grow and differentiate rapidly in a simple saline solution (1), they are also an exceptional system in which to test gene function in developmental processes. This is most often done by microinjecting fertilized eggs. Using a morpholino reagent that inhibits the function of a single target gene, as well as overexpression driven by mRNA injection, can alter the levels of specific target genes. The injected egg can be then grown to various stages from early blastula stages to tadpole stages (within 3 days), then examined for the impact of the microinjection by examining the anatomy or gene expression profile in the tadpole (1). Research on amphibian embryos for over a century (2) has generated a vast and rich literature describing developmental processes, and the explosion in high-throughput sequencing, and *in situ* hybridization is generating an extraordinary resource for understanding the role of genes and gene pathways in basic cellular and developmental processes (3). Xenbase (4,5) aims to incorporate all of these different types of data, to bridge it using ontologies and annotation, in order to allow researchers using *Xenopus* or other model organisms to make basic discoveries relevant to understanding human health and disease.

The Xenbase resource is comprised of two different databases that are seamlessly integrated and generate a single merged view to resource users. The majority of data are stored in a DB2 relational database based on the CHADO schema (6), while data driving the genome

*To whom correspondence should be addressed. Tel: +403 220 8502; Fax: +403 289 9311; Email: pvize@ucalgary.ca
Correspondence may also be addressed to Aaron M. Zorn. Tel: +513 636 3770; Fax: +513 636 4317; Email: aaron.zorn@cchmc.org

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

browsers runs on a MySQL database. This system allows us to utilize GMOD software modules, such as Gbrowse (7) via MySQL with minimal effort, and to use the indexing, power and speed of a relational database for the intense queries necessary for users to retrieve information on complex subjects such as gene expression searches. The data are merged in Apache and served to users.

Xenbase has added many new features since the last report (5) and also vastly increased the amount of manually and automatically curated data. We presently host three genome builds, 51 000 images, 15 844 gene pages, over 4 million nucleotide sequences and over 43 000 publications on *Xenopus*. All of these data are integrated via the *Xenopus* Anatomical Ontology (8) and a synonym matching system that deals with anatomical and molecular term heterogeneity in a seamless manner.

GENOMES AND EPIGENOMIC DATA

The new *Xenopus laevis* genome (version 6.0), is the first public release of the genome of this allotetraploid species. The genome, along with those of *X. tropicalis* discussed below, was generated by the JGI team at the University of California Berkeley. A new build of the *X. tropicalis* genome, version 7.1 along with new gene model predictions (version 7.2) have also been added to the Xenbase Gbrowse implementation.

The Xenbase implementation of Gbrowse (4,7) not only allows users to utilize standard Gbrowse features, it links gene models to all of the relevant data within the main database, such as gene pages, gene expression data, details on ESTs, literature on the gene etc. When a track is clicked a small window opens displaying data on the gene including its symbol and name and images of its expression. A link is provided to directly jump to the corresponding gene page with further information.

Several new tracks are available, including fosmid clones, bac ends, epigenomic data on histone acetylation and methylation, evolutionarily conserved regions, ENSEMBL gene models, Refseq promoters and Affymetrix probes.

NEW AND UPGRADED GENE PAGE TABS

In order to reduce the amount of information displayed on a page, gene data are sorted into a set of tabs. Either an image icon in the case of gene expression images, or a number in brackets for other tabs, indicates that the tab has content. The data in each tab have been expanded and two new tabs added—'Nucleotides' and 'Protein'. The Nucleotide tab illustrates genome-derived predicted gene models and also collates all known cDNA, assembled clusters and ESTs associated with the Gene Page. Data from *laevis*, *tropicalis*, or both, can be viewed. Selecting to rocket icon next to a sequence description will format and load the sequence into our BLAST interface, while the magnifying glass icon will bring the sequence up in a new window for simple copy and paste use.

The Protein tab assembles all associated protein sequence and will soon also contain subcellular localization and post-translational modification data.

ANATOMICAL ONTOLOGY IMPROVEMENT

The *Xenopus* anatomy ontology (XAO) was constructed with a controlled vocabulary representing the anatomy and development of *X. laevis* and *X. tropicalis* (8). It represents not only description of anatomical elements, is also maps the lineage of tissues and the timing of their development. The XAO uses a common framework adopted by several model organism database communities and its topmost nodes are based on the Common Anatomy Reference Ontology (CARO, 9). Ongoing development of the XAO includes adherence to a strict hierarchy of cross-referencing to other bio-ontologies (xrefs). The top-down hierarchy includes CARO, the Uber Anatomy Ontology (UBERON), the Vertebrate Skeletal Anatomy Ontology (VSAO), the Amphibian Anatomy Ontology (AAO) and the XAO (9). Specifically, we continue to include references to the appropriate orthogonal terms present in the next highest level ontology. Recent updates to the XAO included synchronizing terms and definitions with the VSAO and the AAO. This allows the ability to query and communicate across orthogonal biological and human disease databases.

The XAO is integrated with the Xenbase gene expression curation interface as well as the automated population and term linking of Xenopus literature to Xenbase via e-utilities. Manual curation by biocurators directly uses terms from the XAO allowing annotation of gene expression patterns in anatomical structures during stages of Xenopus development. The XAO will also be used to annotate mutant and morphant phenotypes. A phenotype feature on Xenbase is currently being developed. Post-compositional entity-quality (EQ) and entity-quality-entity (EQE) tagging with additional ontologies, such as the Phenotypic Quality Ontology (PATO), Spatial Ontology (BSPO) and the cellular component portion of the Gene Ontology (GO) (10), will aid in curation of phenotypes as well as enhance curation of gene expression patterns during *Xenopus* development.

In this phase of XAO development, we continue to strive for completeness and accuracy for each term contained in the ontology. The previous release of the XAO had 736 anatomical terms, 469 of which lacked definitions. The current version has 1190 terms, all of which are defined. Each anatomical term has also been designated appropriate Nieuwkoop and Faber developmental stages (11). The XAO is now 'is_a complete' meaning each term is assigned one is_a relationship to a supertype. This provides a clean classification structure and a logical framework for 'genus-differentia' type definitions as well as direct child to root organization. We improved the develops_from map we reported earlier (8) and now each term in the XAO now has immediate or inherited develops_from and part_of relationships. For example, the endocardial tube (XAO:0000337) develops_from the endocardium (XAO:0000066) and is part_of the heart (XAO:0000064). Users are now able to take advantage of these relationships to retrieve data for any derivative of neural crest (XAO:0000048) in a single search operation. Using the search feature on Xenbase allows the user to include successor tissues of neural crest of

which include cardiac neural crest from which the heart is derived. We continue to represent synonyms to aid users in anatomy searches using their preferred nomenclature.

The XAO is one of six exemplar ontologies in the Open Biological and Biomedical Ontologies Foundry (9) recommended to serve as targets for community convergence. The XAO, in both OBO and OWL (Web Ontology Language) format, is available for download from the Xenbase FTP (ftp://ftp.xenbase.org) and from OBO Foundry (http://www.obofoundry.org). The XAO can also be browsed under the Anatomy and Development section on Xenbase. We encourage users to send term requests, comments and clarifications, and synonym recommendations via the XAO tracker (https://sourceforge.net/tracker/?group_id=76834&atid=1127722).

ANNOTATION OF GENE EXPRESSION AND LITERATURE ON *XENOPUS*

Xenbase now hosts a compendium of over 43 000 research articles and we have brokered image reproduction agreements with 14 journals in addition to 5 open access journals. Curators use a semi-automated system to append images and figure legends to our literature pages (5). Selected articles are manually annotated for genes referenced, content type (morpholinos, phenotypes, gene expression, RT-PCR, antibodies) and anatomy terms used within the article, adding to the automated curation. Authorship is confirmed and Supplementary Data examined for relevant gene expression images.

Gene expression curation has been the major focus for the Xenbase curation team and new *Xenopus* research is given priority over legacy data published prior to 2000. The gene expression curation interface (Figure 1) is composed of several easy to use modules displaying the image and options for data entry. Curators begin by selecting species (*X. laevis* or *X. tropicalis*) and then annotate gene expression using anatomy terms from the XAO, which are limited by the software to those appropriate for the relevant developmental stage. This enforces correct term usage, often different from that used in publications. Curators also enter and cross-check Genbank accession numbers and/or clone ID. Gene expression curation fills the expression table and is immediately implemented on the live Xenbase site (Figure 2).

We use automated systems whenever possible to facilitate data entry and curation. There are three different avenues used to add content to Xenbase: (i) automated data pipelines that bring in data from external databases; (ii) manual data curation that processes the *Xenopus* scientific literature and community submitted data, then parses it into the database; and (iii) hybrid systems (e.g. the literature image importing system described above) that use elements of both of these approaches to accelerate data entry and maintain quality control. Our data processing pipeline continues to pull in any articles released or updated in PubMed that include the text '*Xenopus*' in the title or abstract, downloading new data weekly. In addition, the curation team routinely checks electronic tables of contents of new journal issues to ensure complete literature coverage. In 2011 Xenbase launched the powerful text-mining tool, Textpresso (12), which processes full text of all newly loaded articles, capturing terms matching various vocabularies—gene symbols, anatomy terms—automatically annotating each article page, and in turn linking out to the gene pages and XAO. Many of our data pipelines utilize NCBI E-utilities. These tools allow us to parse NCBI content directly into Xenbase data tables. A live demonstration of this is available using the 'add new paper' link in the literature section. When a (logged in) user enters a PubMed ID into this field and selects the 'Add Article' button, Xenbase uses E-utilities to query PubMed and download the available XML into the appropriate Xenbase data tables. The literature module then extracts the appropriate content, builds a publication/article page, and sends this back to the user—all within seconds. Over the course of the following week, various automated and semi-automated data processing pipelines associate the addition of new content and mine the papers for content, identify and link gene names and anatomy terms, and attribute the paper authorship to registered Xenbase users. Manual confirmation of authorship builds the publication list for each member of the *Xenopus* research community and clarifies any confusion resulting from common names. Curators manually add any further annotations for content (e.g. genes mentioned but not captured, fine grained anatomy items) to expand literature annotation.

Content at external resources changes constantly, and we have established a number of automated processes to keep our content synchronized. One important example this phenomena is gene nomenclature. Xenbase acts as a clearing-house for *Xenopus* gene nomenclature and assigns gene names and symbols that match Human Gene Nomenclature Committee (HGNC) names. However, some genes do not yet have HGNC names, and names sometimes change as more information becomes available for previously uncharacterized genes. Automated processes match Xenbase gene symbols against HGNC symbols at the NCBI, and when a symbol no longer matches, Xenbase incorporates the new data and updates our symbols and associated gene names. The outdated gene symbol/name is made a synonym for that gene. Another example of dynamic external content are journal publications which have temporary records released upon electronic publication and new records generated when the print version is released (and a third when NCBI annotates the paper). Our system detects the changes and updates the literature records accordingly. Similarly new DNA, mRNA and protein sequences added to the NCBI resource are also identified automatically. These synchronization processes run weekly.

ANTIBODY REPRESENTATION

A new antibody module provides comprehensive support for antibody reagents. This feature was implemented to supplant the existing antibodies wiki page, and it

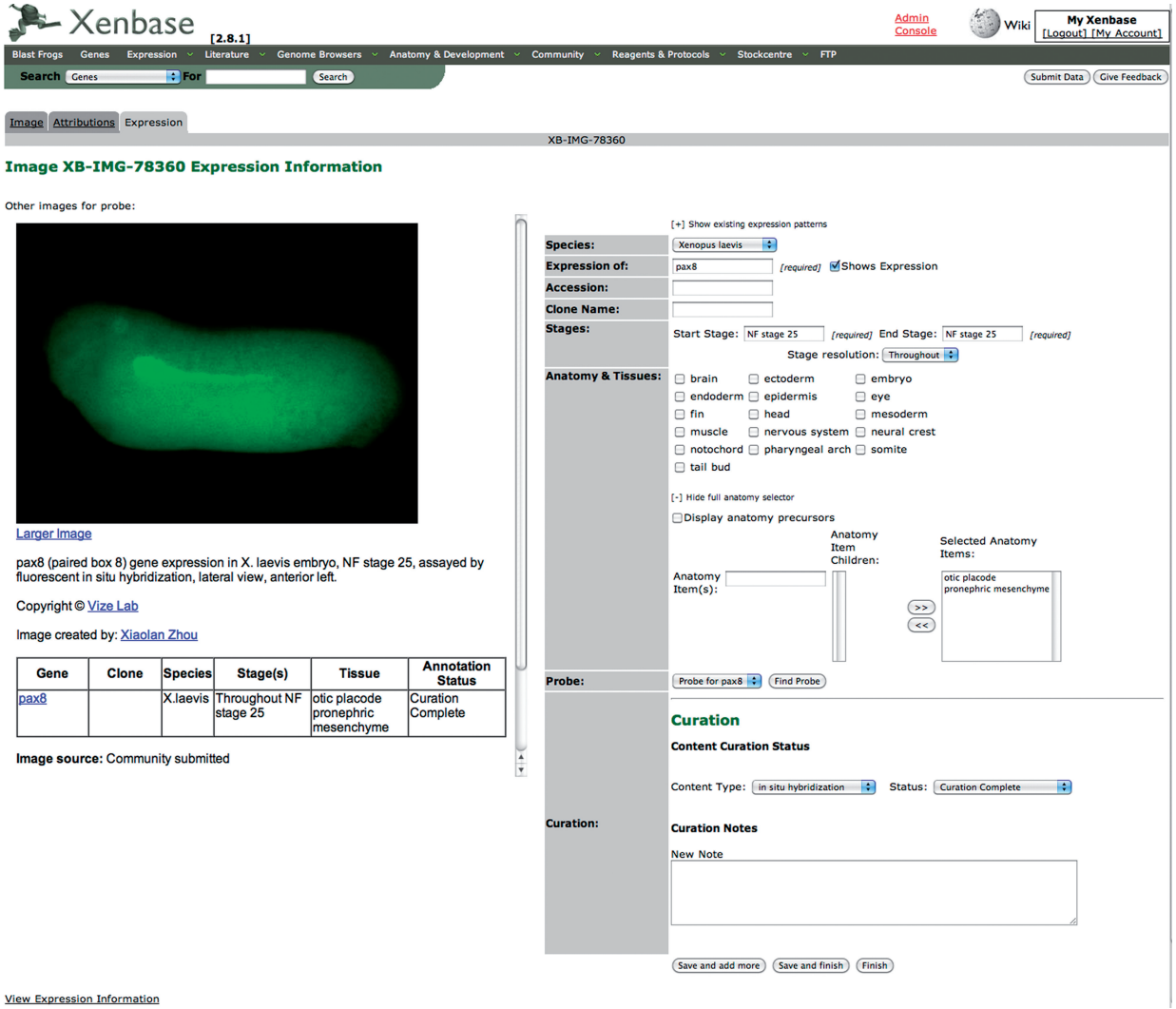

Figure 1. The Xenbase gene expression annotation interface. An image, or panel of images is loaded. After selecting the species and developmental stage, common anatomical terms appropriate for this stage range are loaded. If a term needed, in this example, otic placode, does not appear, the curator types the term in the 'anatomy Item(s)' text field. Once three letters have been typed AJAX generates a list of suggestions that the curator then selects from and the ID appended to the image. Often element names corresponding to adult structures are used by authors, for example kidney, rather than the stage appropriate name for the item at the time it is shown, which in an early embryo may be 'pronephric mesenchyme' as in this example. The system enforces correct stage appropriate term usage.

features enhanced search functions, data entry interfaces and database integration. The wiki contains information on over 700 antibodies gathered from literature and community polling with known reactivity to *Xenopus* antigens. We estimate that the total number of *Xenopus*-reactive antibodies for all experimental applications is in the thousands and are continuously adding them as part of the normal literature curation process. The new interface captures and displays all essential experimental and biological properties of these reagents. Principally, these properties include (i) immunogen (and antigen when known); (ii) gene(s) recognized; (iii) host species and clone isotype; (iv) the experimental application (e.g. western blot or immunohistochemistry); (v) post-translational modification of immunogens; and (vi) working concentrations by application. In addition, the database has the capacity to capture cross-reactivity to multiple genes, tissue or chemical class immunogens such as tissue homogenates or BrdU, and negative reactivity data.

The new antibody feature also allows greater integration with the entire Xenbase database. In contrast to the stand-alone wiki page, antibody records are now indexed to their respective gene pages, and images of embryos processed by immuno-histochemistry with an antibody probe are now available both in publications that released the images and in corresponding gene pages as appropriate. Users can also now browse or search by antibody name, clone, gene, gene synonym, tissue of expression, supplier and experimental application.

GENE EXPRESSION SEARCH UPGRADES

The Gene Expression search has been enhanced to accommodate more complex queries. Users can now find gene expression data associated with a DNA or protein

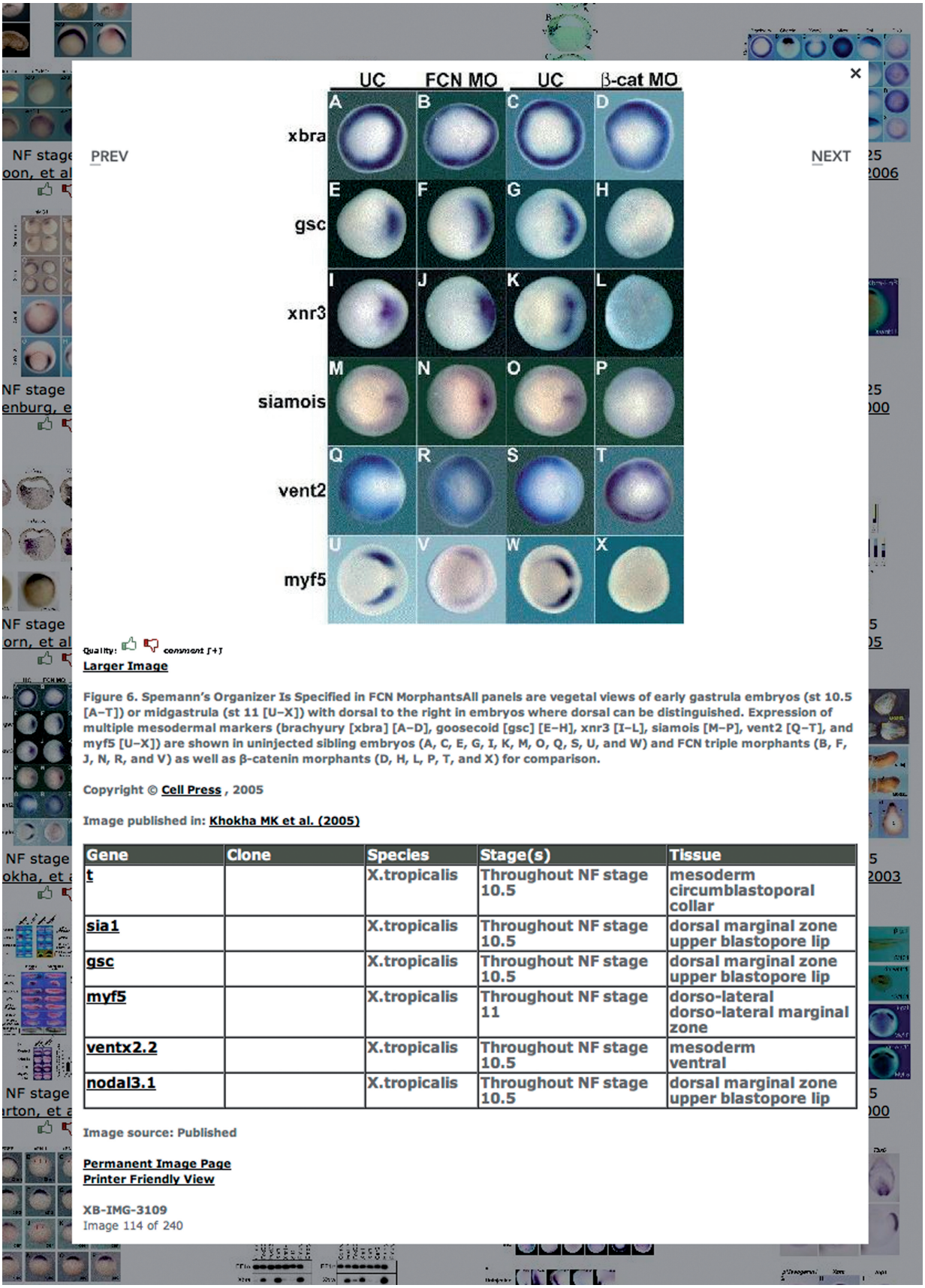

Figure 2. Manual annotation data are displayed along with literature figures and legends. For each image curated from the literature the legend along wit Xenbase annotations are displayed. Often these are quite different due to the use of legacy gene names in the literature or non-standard anatomical terminology. In this example the 't' gene is called 'xbra' in the legend. No detailed anatomical data are present in the figure legend, but have been added by curators and can now be found using the various query systems in the Gene Expression Search toolset.

sequence via BLAST in addition to search by gene name. This approach was first used by Gilchrist *et al.* (13) and avoids the necessity to first find the symbol or name for a gene, although this is very straightforward, once again using BLAST. We have also added an option for searching multiple tissue target with AND or OR combinations, the ability to include successor or predecessor tissues, and the ability to exclude results that also show expression in defined tissues (e.g. expressed in heart and lung but not in skeletal muscle).

Wiki

A wiki has been implemented to allow for community contributions to Gene Pages and as a nexus for posting protocols and reagent data. The wiki stores images of *Xenopus* developmental stages, development time charts at multiple temperatures, movies and images of *Xenopus* development, community announcements and conference notifications and more.

DATA EXPORTS

Most of our data are available in reports that are updated weekly and posted on our ftp site ftp://ftp.xenbase.org/. This includes all nucleotide and protein sequences, gene expression data, gene names and symbols, Affymetrix and Unigene IDs etc. These data are imported and used by a variety of external resources such as NCBI, BeeGee and the Uniprot knowledge base to both supplement their content and to generate reciprocal links. Custom data reports are available on request.

DATA SUBMISSION

Data submissions from the community are strongly encouraged. A 'Submit Data' button is present on every page and takes users to a resource that allows them to upload up to 250 MB of data, along with tools for adding data descriptions etc. Xenbase curators then parse this material into the appropriate tables as appropriate. Larger data submissions can be achieved via our large-scale ftp data storage site, which has a multi-terabyte capacity.

CITING XENBASE

If Xenbase content or services contribute significantly to a publication or need citation, users are suggested to cite this article.

FUTURE DIRECTIONS

Xenbase now stores over 4 GB of data in many hundreds of tables. In the next few years we will work towards implementing developmental phenotypes. Not only will Xenbase represent how anatomical structures are altered in response to changes in gene expression, we will also record how gene expression in downstream genes is impacted when a target gene is up- or down-regulated. In *Xenopus* experiments often one target is down-regulated by microinjecting a morpholino, and the impact is assayed by whole-mount *in situ* hybridization using a bank of marker genes (14). Other goals are upgrades to the genome browser and displaying quantitative RNA-seq data in Gbrowse, a large-scale data repository and the high-quality annotation of both *laevis* and *tropicalis* genomes.

FUNDING

Xenbase is supported by grants from the Eunice Kennedy Shriver National Institute of Child Health and Human Development [award numbers R01HD045776 and P41HD064556 to A.M.Z. and P.D.V.]. Funding for open access charge: NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Sive,H.L., Grainger,R.M. and Harland,R.M. (2000) *Early Development of Xenopus laevis*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
2. Roux,W. (1885) Uber die bestimmung der hauptrichtungen des froschembryo im ei und uber die erste theilung des froscheies. *Zeitschrift*, **20**, 1–54.
3. Grainger,R.M. (2012) *Xenopus tropicalis* as a model organism for genetics and genomics: past, present, and future. *Methods Mol. Biol.*, **917**, 3–15.
4. Bowes,J.B., Snyder,K.A., Segerdell,E., Gibb,R., Jarabek,C., Noumen,E., Pollet,N. and Vize,P.D. (2008) Xenbase: a *Xenopus* biology and genomics resource. *Nucleic Acids Res.*, **36**, D761–D767.
5. Bowes,J.B., Snyder,K.A., Segerdell,E., Jarabek,C.J., Azam,K., Zorn,A.M. and Vize,P.D. (2010) Xenbase: gene expression and improved integration. *Nucleic Acids Res.*, **38**, D607–D612.
6. Mungall,C.J., Emmert,D.B. and FlyBase Consortium. (2007) A chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.
7. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
8. Segerdell,E., Bowes,J.B., Pollet,N. and Vize,P.D. (2008) An ontology for *Xenopus* anatomy and development. *BMC Dev. Biol.*, **8**, 92.
9. Noy,N.F., Shah,N.H., Whetzel,P.L., Dai,B., Dorf,M., Griffith,N., Jonquet,C., Rubin,D.L., Storey,M.A., Chute,C.G. *et al.* (2009) BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.*, **37**, W170–W173.
10. Washington,N.L., Haendel,M.A., Mungall,C.J., Ashburner,M., Westerfield,M. and Lewis,S.E. (2009) Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol.*, **7**, e1000247.
11. Nieuwkoop,P.D. and Faber,J. (1994) *Normal Table of Xenopus Laevis (Daudin)*. Garland, New York.
12. Muller,H.M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, **2**, e309.
13. Gilchrist,M.J., Christensen,M.B., Harland,R., Pollet,N., Smith,J.C., Ueno,N. and Papalopulu,N. (2008) Evading the annotation bottleneck: using sequence similarity to search non-sequence gene data. *BMC Bioinform.*, **9**, 442.
14. Kenny,A.P., Rankin,S.A., Allbee,A.W., Prewitt,A.R., Zhang,Z., Tabangin,M.E., Shifley,E.T., Louza,M.P. and Zorn,A.M. (2012) Sizzled-tolloid interactions maintain foregut progenitors by regulating fibronectin-dependent BMP signaling. *Dev. Cell.*, **23**, 292–304.